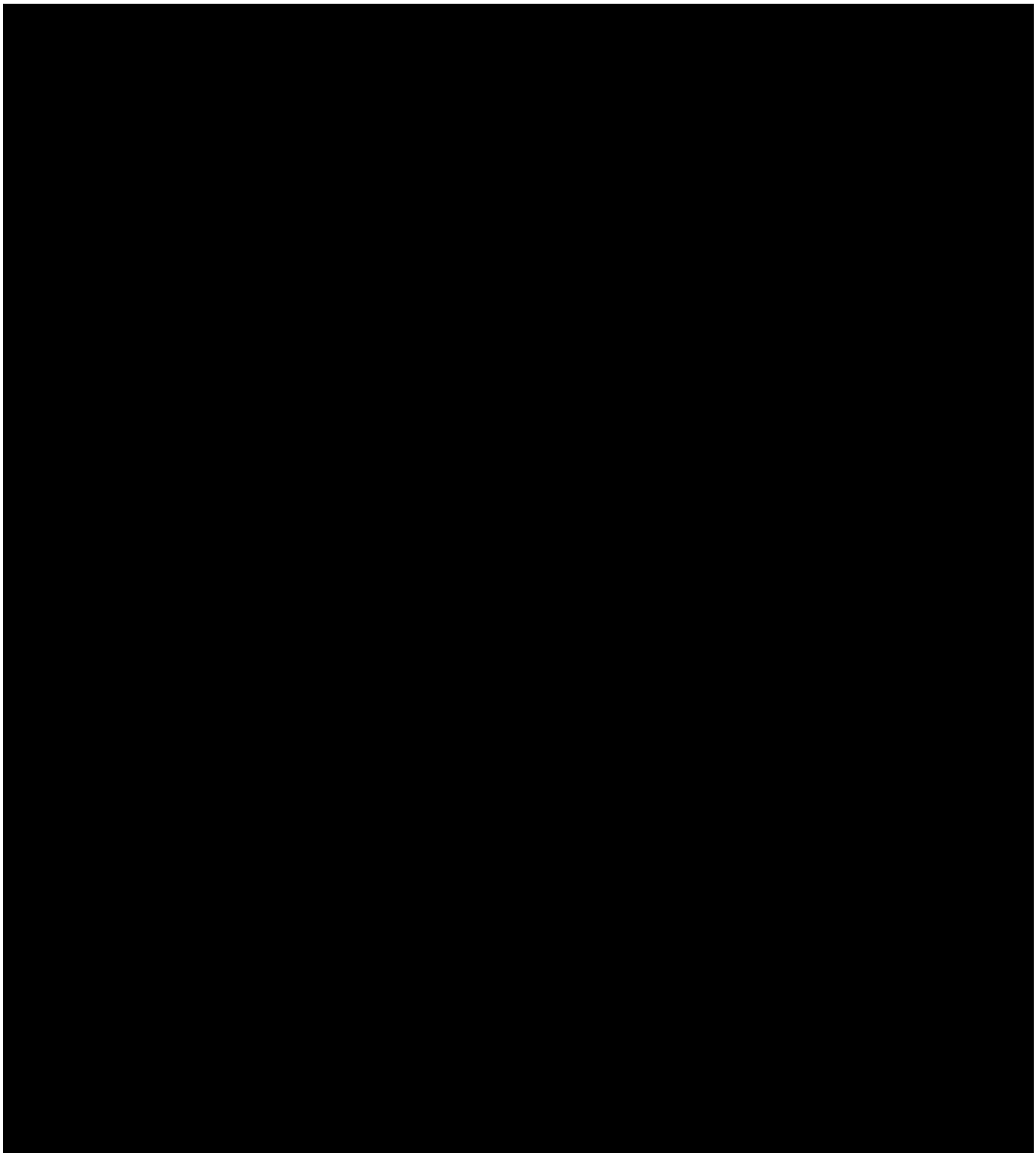
EXHIBIT 5

UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC, Plaintiff, v. GOOGLE LLC, Defendant.	Civil Action No. 1:19-cv-12551-FDS Hon. F. Dennis Saylor IV

EXPERT REPORT OF SUNIL P. KHATRI, PhD.

[redacted]



[0062] As the patents-in-suit explain, Dr. Bates recognized that even though then-modern conventional microprocessors contained about one billion transistors, they could perform only "a few operations per

clock cycle."[2] '156 patent, 3:12-14,1:55-63. Dr. Bates explained that a large portion of this inefficiency comes from using transistor-intensive full-precision arithmetic units:

> As described above, today's CPU chips make inefficient use of their transistors. …they deliver great precision, performing exact arithmetic … with 32 and 64 bit floating point numbers. Many applications need this kind of precision. As a result, conventional CPUs typically are designed to provide such precision, using on the order of a million transistors to implement the arithmetic operations.

'156 patent, 3:11-3:26.

[0063] However, Dr. Bates realized that such full-precision, inefficient components were not necessary for all applications, including many valuable ones:

> There are many economically important applications, however, which are not especially sensitive to precision and that would greatly benefit, in the form of application performance per transistor, from the ability to draw upon a far greater fraction of the computing power inherent in those million transistors. Current architectures for general purpose computing fail to deliver this power.

'156 patent, 3:27-33.

[0064] The patents-in-suit are thus directed away from prior art computers based on full-precision execution units that take up space and are wasteful of transistors.

[0065] As Dr. Bates further explains in the specifications of the patents-in-suit, "[b]ecause LPHDR processing elements are relatively small, a single processor or other device may include a very large number of LPHDR processing elements, adapted to operate in parallel with each other." '156 patent, 6:56-59. As a result, "embodiments of the present invention may be implemented as any kind of machine which uses LPHDR arithmetic processing elements to provide computing using a small amount of resources (e.g., transistors or volume) compared with traditional architectures." '156 patent, 8:8-12.

[0066] By using a "very large number" of LPHDR execution units in parallel, computer systems are able to achieve significantly better performance than prior art systems. Because each LPHDR execution unit requires fewer resources (*e.g.*, fewer transistors, less physical volume) than a full-precision execution

[2] In this report, I will frequently refer to a "cycle" or a "clock cycle" in explaining my opinions. Broadly speaking, computer circuitry operates at a pace that is dictated by a "clock" circuit, which generates regular edges (or "ticks") that allow different circuits to synchronize and coordinate their actions. The duration of a clock cycle is the period of time between one clock rising (or falling) edge and the next; it is – in effect – the time it takes to perform the smallest operation in a computer system. Some operations can be performed in a single clock cycle, while more complicated computing operations may require several clock cycles to complete.

In modern computer architectures, it may take multiple clock cycles to perform an arithmetic operation. In such cases, it is common to design execution units using circuits that are organized in a "pipelined" fashion, with the first stage of the pipeline performing the first step of one operation while the next stage of the pipeline performs the second step of the operation, and so on. *See, e.g.*, GOOG-SING-00022834 at 3204-3223 (chapter describing "Pipelined Datapath and Control" from *Computer Organization and Design*," by Patterson and Hennessy) . With pipelining, the system can complete one operation every clock cycle (in this case, we say that the throughput of the system is one clock cycle), even though it may take multiple clock cycles between the time a specific operation is issued to the pipeline first stage, and completed by the last pipeline stage (in this case, we say that the latency of the system is *n* clock cycles, where there are *n* pipeline stages). The efficiency and/or speed of a computer system is often measured not by how many clock cycles any particular operation requires to complete (latency), but rather by the average number of operations that can be completed (its throughput) per clock cycle (or per second, when comparing computers with different clock speeds). One common such metric is "FLOPS", which stands for "Floating-Point Operations Per Second." *See, e.g.*, GOOG-SING-00022834 at 2930 ("Another common performance figure is MFLOPS (million of floating-point operations per second)").

[redacted]

unit, "there is a large amount of arithmetic computational power per unit of resource. This enables larger problems to be solved with a given amount of resource than does traditional computer designs." '156 patent, 23:37-44; *see also id.*, 6:56-60. In particular, the claimed systems "might perform tens of thousands of arithmetic operations per cycle, as opposed to hundreds in a conventional GPU or a handful in a conventional multicore CPU. '156 patent, 23:46-49.

[0067] In addition, the patents-in-suit also teach computer systems in which the number of LPHDR execution units exceeds the number of full precision execution units:

> For certain devices … according to the present invention, the number of LPHDR arithmetic elements in the device (e.g., computer or processor or other device) exceeds the number, possibly zero, of arithmetic elements in the device which are designed to perform high dynamic range arithmetic of traditional precision (that is, floating point arithmetic with a word length of 32 or more bits).

'156 Patent, 27:52-59.

[0068] The increased level of compute parallelism and scale in such computer systems is necessarily achieved at the cost of precision—the vast majority of the high dynamic range floating-point operations performed by the device must be performed at low precision. Dr. Bates was the first to understand that sacrificing precision for increased parallelism/scale results in significant performance gains per unit of resource over the prior art. In fact, Dr. Bates notes that when certain applications are implemented using a device that uses LPHDR execution units, the final application error is significantly lower than the error of the LPHDR execution units themselves. '156 patent, 16:59-23:34.

[redacted]

[redacted]

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Date: December 22, 2022.

Sunil P Khatri, Ph.D.

[redacted]